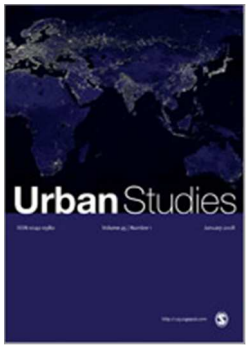

Understanding urban gentrification through Machine Learning

Journal:	*Urban Studies*
Manuscript ID	CUS-727-17-08.R2
Manuscript Type:	Article
<b>Discipline: Please select a keyword from the following list that best describes the discipline used in your paper.:	Geography
World Region: Please select the region(s) that best reflect the focus of your paper. Names of individual countries, cities & economic groupings should appear in the title where appropriate.:	Europe
Major Topic: Please identify up to 5 topics that best identify the subject of your article.:	Displacement/Gentrification, Method, Neighbourhood, Redevelopment/Regeneration, Housing
You may add up to 2 further relevant keywords of your choosing below::	Machine Learning, Geographic Data Science

Understanding urban gentrification through Machine Learning: Predicting neighbourhood change in London

Abstract

Recent developments in the field of machine learning offer new ways of modelling complex socio-spatial processes, allowing us to make predictions about how and where they might manifest in the future. Drawing on earlier empirical and theoretical attempts to understand gentrification and urban change, this paper shows it is possible to analyse existing patterns and processes of neighbourhood change to identify areas likely to experience change in the future. This is evidenced through an analysis of socio-economic transition in London neighbourhoods (based on 2001 and 2011 Census variables) which is used to predict those areas most likely to demonstrate 'uplift' or 'decline' by 2021. The paper concludes with a discussion of the implications of such modelling for the understanding of gentrification processes, noting that if qualitative work on gentrification and neighbourhood change is to offer more than a rigorous post-mortem then intensive, qualitative case studies *must* be confronted with—and complemented by—predictions stemming from other, more extensive approaches. As a demonstration of the capabilities of Machine Learning, this paper underlines the continuing value of quantitative approaches in understanding complex urban processes such as gentrification.

Keywords

London, neighbourhood change, gentrification, principal components, machine learning, random forests, census, quantitative geography

Introduction

The application of quantitative methods to the study of neighbourhood change in general—and gentrification in particular—still has something of a controversial air. Despite some of the most-cited works in the field utilising quantitative methods to either measure the 'rent gap' between actual and potential housing rents (*e.g.* Ley 1986; Clark 1988) or demonstrate socioeconomic change through census analysis (*e.g.* Atkinson 2000; Hamnett 2003), the majority of literature on gentrification now shuns quantitative analysis in favour of qualitative assessments of neighbourhood change based on media analysis, interviews, ethnography and other forms of observational data collection. In part, this is because of the limitations of secondary data for capturing the dynamics of urban processes occurring at a local level (Watt 2008), but this is often coupled with a suspicion that 'official' statistics relating to neighbourhood change describe patterns but obfuscate underlying processes of class change (Slater 2009).

Consequently, in most contemporary accounts, intensive and qualitative methods are the favoured means of exploring urban gentrification; however, the privileging of such methods is not without risks since, as Barton (2016: 92) points out, "qualitative strategies for identifying gentrified neighbourhoods may overlook areas that experienced similar changes to those more widely recognised as gentrified." Focusing on New York, Barton (2016) and others (*e.g.* Bostic and Martin, 2003; Freeman, 2005) use regression methods to reveal a much larger number of census tracts where gentrification seems to have occurred than those generally highlighted in the literature. This suggests that the academic and media preoccupation with Brooklyn and Manhattan districts experiencing obvious social and cultural change (*e.g.* a transition from black to white occupation and the associated rise of 'hipster' stores) distracts from a wider appreciation of the situation across the five Boroughs.

In other cities, a similar privileging of select 'signifying locations' appears equally evident, with certain neighbourhoods repeatedly attracting the researcher's gaze; as Neal *et al*.

(2016) wittily put it: 'You can't move in Hackney without bumping into an anthropologist'. Indeed, recent analyses of London have fixated on specific parts of the East End (*e.g*. Harris 2012 on Hoxton; Watt 2008 on Stratford; and Butler *et al*. 2013 on Hackney) or South London (*e.g*. Jackson and Benson 2014 on Peckham; Mavromattis 2012 on Brixton), potentially ignoring other neighbourhoods where significant change is occurring. Quantitative and multivariate analysis across a range of neighbourhoods hence appears important for grasping the bigger picture and, more importantly, it appears such methods could predict where the 'gentrification frontier' might move to next (see Chapple 2009).

The work presented here provides a quantitative analysis of this kind and is motivated by the emergence of 'machine learning' techniques (hereafter: ML) that have the capacity to learn from, and make predictions about, observations in large data sets without being explicitly programmed with a model of how to do so. We will detail our specific approach later, but suffice to say here that most ML approaches incorporate some form of optimisation (a measure of whether the predictions are getting better or worse), alongside phases of training (in which the algorithm learns how to make predictions based on 'existing' data) and testing (in which results are tested for robustness using 'new' data).

While such methods will not necessarily lead to new theories of gentrification on their own, in this paper we suggest that they can indicate possible *trajectories of neighbourhood change*, something that is particularly important in theory development (Owens 2012). We explore this contention by using the 'random forests' algorithm to tease out the trajectories of 4,835 London neighbourhoods between 2001 and 2021, based on analysis of social, economic and environmental variables. The contribution of this paper to gentrification debates is not, however, solely methodological (*i.e*. showing how we can use ML methods to predict urban change) but also empirical (*i.e*. mapping shifts in London's 'gentrification frontier' via a fine-grained analysis of neighbourhood change).

Modelling neighbourhood change

It has been suggested that gentrification needs to be understood as a neighbourhood-level phenomenon involving not just an increase in the value of an individual property, but a simultaneous uplift in the values of comparable properties across a given neighbourhood (O'Sullivan 2002). In classic theories of gentrification this uplift is associated with the arrival of new, wealthier populations and the displacement of existing inhabitants, alongside improvements to the housing stock that register this socio-economic transition (Atkinson 2000). Alternative theories suggest that improvements to the built environment can also occur via *marginal gentrification* caused by the arrival of culturally-rich—though not necessarily affluent—populations, such as artists and students (Hochstenbach *et al*. 2015), and via *incumbent upgrading* by longer-term residents (Van Criekingen and Decroly 2003). Owens (2012:347) operationalised these in a quantitative context using the concept of neighbourhood 'Socio-Economic Status' (SES) change: we adopt this given it potentially reveals change-processes other than gentrification and displacement *per se*.

Notwithstanding the risk that some neighbourhood processes occur at a granular level that cannot be 'seen' through quantitative data (Barton, 2016: 99), there remains the challenge of defining a neighbourhood in the first place. Here, there are a host of overlapping definitions available, but for our purposes the one advanced by Galster (2001: 2112) offers a suitable starting point: "the bundle of spatially-based attributes associated with clusters of residences, sometimes in conjunction with other land uses." While this does not establish neighbourhoods as discrete, bounded entities (*i.e*. it does not unambiguously state how big or small a neighbourhood is), it provides a basis for defining neighbourhoods on different spatial scales through the 'bundling' of attributes. In effect, Galster defines a set of 'domains' within which neighbourhood-ness is constructed, namely: urban morphology; mobility and utility infrastructures; demography; class; tax and public services; the environment; proximity to facilities (both recreational and employment-based); political networks; degree of social interaction; and sentiment (*i.e.* place attachment).

In a US context, Van Criekingen and Decroly (2003:2457) employed indicators of deprivation, upgrading of the built environment, social status, population, and income change to classify neighbourhoods on this basis. Here, there are obvious parallels to geodemographic analyses of the type underpinning the operationalisation of the 2001 and 2011 Output Area Classifications in the UK (Vickers and Rees 2007; Gale and Longley 2011; Gale 2014; and see also Li and Xie 2018 on the clustering of US census data, 1970–2010). But while geodemographics uses area attributes to assign neighbourhoods to groups (*i.e*. clusters), we use these attributes to predict an outcome.

Contextualising machine learning in urban studies

To date, ML has most commonly been employed in physical geography where it is often used in conjunction with remotely-sensed data to classify landforms (Xiao 2016). Recently, the use of ML in topics of interest to human geographers—such as changes to the fabric of cities, the prediction of transport modality, detection of deprivation, and population prediction—has grown rapidly as well (*e.g.* Arribas-Bel *et al*. 2011; Arribas-Bel *et al*. 2017; Donaldson and Storeygard 2016; Hagenauer and Helbich 2017; Naik *et al.* 2017; Liu *et al*. 2017; Santibanez et al. 2015; Stevens *et al*. 2015). Revisions to classical regression techniques have also yielded geographically-aware ML tools such as Spatially-Filtered Ridge Regression (Fan *et al*. 2016), and derived probability transitions aiding understanding of the evolution of regional income disparities (Rey 2014).

Because ML differs radically from approaches commonly employed by social science researchers it is worth clarifying what ML can—and cannot—accomplish. The most obvious difference to conventional methods is simply one of scale: ML algorithms not only tackle very ‘long’ data sets containing many rows, they also tackle very ‘wide’ ones incorporating many correlated variables (as intercorrelation does not impact ML approaches in the same way as traditional multivariate analysis, meaning methods can make better use of the full extent of the data). Clearly, a not coincidental reason for the rise of ML is the growing availability of ‘big data’ about human society: telephone usage

(Reades and Smith 2014), vehicle licensing (Lansley 2016), public transit smartcard usage (Zhong *et al*. 2014), and even taxi trips (Manley *et al*. 2015) are all amenable to analysis. Of course, many cultural aspects remain 'off the radar' (Barton 2016:94), but in the context of neighbourhood change, social media such as Twitter or Instagram, and even Tripadvisor reviews, can offer useful proxies (see Boy and Uitermark 2016; Hristova *et al*. 2016; Zukin *et al*. 2017).

Unlike conventional statistical methods, ML approaches are not necessarily concerned with causality, being primarily concerned with *utility*. The online retailer Amazon, for instance, does not care why there is a strong relationship between two books in its customers' purchasing patterns, only whether they can influence the customer to buy the second book. As Wyly (2014:681) puts it: "The capitalist correlation imperative is clear: spurious correlation is fine, so long as it is *profitable* spurious correlation." The capacity of modern corporations to 'consume' large volumes of data with which to make profitable predictions is *one* outcome of the rise of ML and 'big data', but the availability and openness of these tools—they are not 'black boxes' to quite the extent that Dalton and Thatcher (2015) appear to believe—means that researchers are now in a position to create 'early warning systems' (Chapple 2009; Chapple and Zuk 2016; Steif *et al*. 2017) to alert residents, representatives, and policy-makers to incipient changes in an area's social and economic dynamics.

This noted, the research undertaken in this article explores neighbourhood change in London using 166 variables across transport, housing, demographics, income and wealth, amenity, and occupational domains. Ultimately, this article does not seek to provide new insights into the root causes of gentrification—these have been amply covered elsewhere in the literature (*e.g*. Davidson and Lees 2005; Hamnett 1984; Redfern 1997, 2003; Zukin *et al*. 2009)—but uses contemporary ML techniques to help select features (*i.e.* variables) from the available data in one time period that might be useful for predicting status change in the next, and to use the outputs of our model to foster debate about the changing urban geographies of the Greater London Authority (which includes 32 London Boroughs and the City of London).

Methodology

As we noted above, with the principal exception of work by Hamnett (1983, 2003, 2009, 2015), census data has been sparingly used in studies of gentrification and neighbourhood change in the UK. In contrast, North American studies have more frequently used secondary data (*e.g*. Barton 2016; Bostic and Martin 2003; Freeman 2005, 2009; Meligrana and Skaburskis 2005; Owens 2012). In one early study, Melchert and Naroff (1987:681) employed logistic regression on data for Boston, MA to establish that 'amenity, social, housing and economic variables [*have*] predictive capabilities [*that are*] quite substantial… [*indicating*] that the general context of a neighbourhood is of far greater significance than individual groups of characteristics.'

The utility of regression may, however, be severely impacted by collinearity (such as might be expected between education and income, or income and property prices). This inter-dependence is often associated with instability in the model thanks to the 'inflation' of coefficients such that some inputs gain in significance at the expense of other, equally important but partially correlated, variables. Stepwise regression was an early computational means of trying to cope with this challenge, but has now been superseded by more robust approaches—generically and collectively referred to as ML—and it is for this reason that this paper explores the potential of ML for advancing understanding of neighbourhood change.

There are obvious limits to how fully we can document our method, so we focus here on the key steps. However, an important overarching consideration is the importance of open, replicable research (*e.g*. Singleton *et al.* 2016); by using both open data and open source code, we enable replication (Brunsdon 2016) by researchers, activists, policymakers, or even real-estate developers. Indeed, our analysis employs only open data (from the 2001 and 2011 UK Census of Population and the London Data Store - an extensive open data portal). Any reader who disagrees with our methodological choices is also free to adapt the code since this is also freely available—for downloading, revision, and (re)running—as a series of Python-based 'notebooks' on the GitHub code-sharing web site.

Data Assembly

A predictive model of neighbourhood change needs two sets of variables: those that measure the status of a neighbourhood, and those that help us predict changes to come. But even before we get to variable selection, it should be noted that the quantitative analysis of neighbourhoods presents several practical challenges, not least of which is the selection of an appropriate geographical scale. Lauria and Stout (1995) have argued that a block-by-block analysis is essential, but cutting against this claim are two inter-related issues: firstly, that fine-scale data are often considered highly sensitive and suppressed from census outputs; and, secondly, that natural variation between smaller areas yields statistically significant—but not actually meaningful—fluctuation (*i.e*. noise). A good example of the latter would be property prices: at the *street* level, the 'average house price' in any given year might be based on a single transaction for an unrepresentative property! Conversely, larger areas generally lack a sense of cohesion and shared identity that we might associate with a similar quality of life, housing conditions, access to services and so on, and necessarily tend to smooth out variation to undermine the detection of change.

Putting these contradictory effects together suggests it is easiest to work with intermediate or meso-scale data; fortunately, the Office for National Statistics (ONS) provides one such grouping in the Lower Layer Super Output Area (LSOA) (broadly similar to a US census tract). The LSOA contains between 1,000 and 3,000 inhabitants living in between 400 to 1,200 households: a geography small enough that even modest changes in the makeup of an area should show up, but large enough that the sample size of each is statistically robust. Whilst data is available at both finer (*e.g*. Output Areas) and coarser scales (*e.g*. wards or Middle Layer Super Output Area), work in the UK concludes that LSOAs exemplify the characteristics of spatial proximity and social homogeneity which are revealing of "neighbourhood effects" (van Ham *et al.* 2012).

So although LSOAs are statistical units rather than an empirical reality, they are broadly coterminous with the kinds of environments that appear important in giving residents both a sense of identity and a context for everyday life. In fact, up to a point LSOAs are

deliberately constructed to contain a broadly-consistent housing type and demography (see Cockings *et al*. 2011). Analysis at this scale hence provides the main basis for understanding the production of neighbourhoods as socially meaningful and physically distinctive urban spaces in London (Sturgis *et al*. 2014).

Calculating Scores

If we begin by assuming that the indicators identified by Van Criekingen and Decroly (2003) are sufficiently comprehensive then—drawing on Owens (2012)—we can use four variables to measure neighbourhood status: household income (using the modelled median value in each neighbourhood[1]), property sale value (also using the median value), occupational share (the percentage of the neighbourhood's residents in the 'top' occupational classes), and qualifications (the percentage of residents achieving NVQ Level 4 or above). Though private sector rents would have been a useful complement to sales, historical data for this domain is very limited in the UK.

To train the ML algorithm to predict neighbourhood change we need to combine these four variables into a singular measure of 'socioeconomic status'. Since we are working with a long but fairly-narrow data matrix, Principal Components Analysis (PCA) is an obvious choice as it will yield just four components: by taking just the first one we capture the majority of the variation in the input data using a single numeric value. This will necessarily cause *some* loss of detail about neighbourhoods because we do not retain any of the subsidiary components, but we can quantify this loss using the percentage of variance explained by each component (this is also the approach taken by Owens, 2012, following

[1] Household income is not normally available at the LSOA scale in Britain, but the Greater London Authority undertook a modelling project incorporating access to restricted data to produce this for London.

Morenoff and Tienda, 1997). Additionally, we apply PCA simultaneously to both census years to avoid the problem that scores for different years are not directly comparable.

The construction of these scores necessarily entailed decisions about the re-scaling of variables since differences in magnitude could allow one dimension to dominate (*e.g.* house prices vs. share of high-qualifications). Simple unit scaling (*i.e.* remapping the range of each variable to the scale 0–1) is unlikely to address this problem because the existence of 'heavy tails' would lead to the bunching of the data at one end of the scale. Equally, since house prices and incomes are also highly-skewed, the mean is unlikely to be a robust measure of centrality. Robust standardisation using the median and Inter-Quartile Range (IQR) addresses both issues: it preserves outliers while producing comparable scales for the bulk of the data. In our testing, this approach yields the most consistent performance and was applied to all score dimensions. More aggressive, non-linear transformations are possible for extreme distributions prior to this step, but these typically lead to the loss of information about the magnitude of outliers or the balance between dimensions in the score.[2] To ensure that the two census years are directly comparable we apply the same transformation to both.

Selecting Predictor Variables

In line with previous work in this area we attempted to select variables from a range of categories including: Housing, Households, Work, Travel and Amenity. This set is far from exhaustive, and the use of more built environment and amenity features (*e.g.* schools) would be one obvious areas for improvement; however, these nonetheless encompass the principal areas on which work on gentrification and neighbourhood change have focussed. Rather than reproduce the full list of 166 variables, readers are invited to access the additional details in the online repository. Of course, the alert reader will have realised that

[2] The code on GitHub also allows readers to apply Box-Cox and Log transformations to these data to explore the impact of scoring changes on the overall results.

some variables will necessarily play a role in both scoring *and* prediction so it is inevitable that the scores will be correlated with property price, income, skills, and occupation data.

Relative vs Absolute Measures

Lees (2000:403) argues both 'contextuality and scale are significant' in gentrification research, implying the need to incorporate *relative* measures of change as part of any neighbourhood analysis. For instance, given trends in London it is entirely conceivable that an area can experience 'ascent' (*i.e*. an absolute 'improvement' in its score) but at a lower rate than its neighbours (*i.e*. a relative 'decline'). Equally, if gentrification is understood in terms of in-movers having a multiple of the current residents' median income, then 'super-gentrification' (Lees 2003; Butler and Lees 2006) may appear quite similar to 'plain old' gentrification in a relative sense. This is a 'feature' and not a 'bug' of this approach: we can use relative to change to effectively classify *both* as forms of gentrification even if they differ in an absolute sense.

On a practical note, raw values can also be problematic for ML because 'decision boundaries'—the thresholds used for regression or classification—will almost certainly shift over time. For instance, if crime generally falls across London between 2001 and 2011 then a 'low' rate of neighbourhood crime in one Census year is *not* the same as a low rate in the next Census year. Consequently, judged in absolute terms many more areas will appear to have become attractive to gentrifiers even if the relative differences between areas remain substantial. Similarly, even if the relative proportions for each demographic group in city remain the same, an expansion in the absolute number of households could lead to housing stress if supply fails to keep up with demand (Hamnett 2015:244).

Random Forests

Random Forests (see James *et al.* 2013 for a systematic introduction) are a particularly versatile and robust form of non-parametric ML, able to perform both classification (assigning observations to classes) and regression (predicting values from observations) tasks quickly, without much tuning and with minimal bias (Breiman 2001). The term 'random' originates from the way that Random Forests (RFs) employ random subsets of the

available dimensions (*i.e.* variables) to avoid the risk of over-fitting. RFs are *ensemble* methods, meaning they aggregate the output of a large number of *decision trees*—many trees yields one forest—and so can cope with complex, non-linear decision boundaries. We tackle this terminology and its import below.

To understand more fully how this approach works, let us take a simple decision tree: anyone who has played the game Twenty Questions has employed a decision tree since, with each new question, the player divides the 'answer space' into two smaller spaces, one of which is excluded from subsequent consideration (*e.g.* is it bigger than a shoebox? Is it alive?). Shallow trees employing a relatively short sequence of questions can uniquely identify a single 'thing' from a very large number of possible 'things' remarkably quickly. Twenty Questions is a classification problem, but this approach can *also* be used for regression: is it before 10am? After 8am? Is it a weekday? A highway? Applying these questions to some movement data we can predict rush hour volumes. James *et al*. (2013:306) describe the function of a tree as 'prediction via the stratification of the feature space' using a two-step process: the predictor space is divided into a set of 'distinct and non-overlapping regions' and for every observation falling into a given region we make the same prediction (usually the mean of observations from the data used when growing the tree). We will unpack this statement later, but by way of an illustration we show in **Error! Reference source not found.** *part* of an actual tree—one of the many grown by the Random Forest on the data—created as part of this research.

[Insert Figure 1 here]

Although trees can be manually created using expert knowledge, their growth can also be automated using a 'heuristic': typically, the computer selects the dimension that best-enables it to split the data set into two dissimilar groups. At each 'node' (branch in the tree) we deal with progressively smaller subsets of the data and this process continues down each

branch until some stopping point—termed a 'leaf'—is reached. The RF grows each tree on a *randomly*-selected subset (*S*) of all dimensions (*D*); these subsets overlap such that trees use similar, but not identical, subsets of *D*. Randomness is then used *a second time* since the tree is further restricted to considering a random subset of *S* with which to split the 'remaining' data at each node. This approach decorrelates the trees by preventing an over-reliance on any one variable and so helps to prevent over-fitting of the data.

The many trees in the forest then 'vote' as an *ensemble* on their preferred class or predicted value, but the poorly-performing trees tend to cancel each other out (noise) while the useful ones (signal) carry the day. In fact, our model goes further than this by employing the computationally-efficient 'extremely randomised trees' (Guerts *et al.* 2016): this not only employs randomly selected dimensions, it also uses random 'cut points' for each split. The prominence of randomness in this method might seem strange to some readers, but in statistical terms it is highly robust.

Training & Testing

An important component of most ML approaches is the incorporation of training and testing regimes: we train the algorithm on a random subset of the full data set, and then test its performance against the portion of the data set not already used. *K*-fold cross-validation is a common approach: the full data set is split into *k* 'folds', each of which is used *k-1* times as part of the training data set, and *once* as the testing data to be predicted. This has a significant impact on the model's overall bias and helps to ensure that outliers do not unduly impact the model. Here, randomisation again helps improve the robustness of our predictions.

Hyperparameter Tuning

Finally, and in common with many ML approaches, we still need to define how the algorithm should 'learn' about the data and gauge its performance. The RFs learning process is governed by 'hyperparameters' and the most important considerations are:

- That more estimators (trees) may yield more nuanced predictions but can overfit some data.
- That trees can be grown to any depth, but specifying a maximum depth reduces the risk of overfitting with 'deep' trees.
- That the minimum size of leaves should normally be a small number (higher resolution predictions) but can also lead to overfitting with some data.
- That reducing the proportion of features used by a tree helps to manage correlation with other trees by reducing their overlap.

Together, these hyperparameters constitute a 'space' that can be systematically explored as part of the model configuration process. We divide this space into a grid and test every combination of hyperparameters using the *k*-fold training approach set out above. We can compare the performance of each configuration using the Mean Squared Error or Mean Absolute Error of the predictions. It is also possible to generate a R^2 value, although using this metric for direct model comparison is considered problematic.

Neighbourhood change in London 2001-11

To recap, we are using a model built on the characteristics of LSOAs from the 2001 Census to 'predict' the 2011 scores, and then use same model with the 2011 Census data to predict outcomes in 2021. Obviously, predictions remain extrapolations (however sophisticated), and predicting the future is always fraught with difficulty: Hamnett (2003) expected that Clapton in East London would prove resistant to gentrification but it is an area that is now very much on—or even behind—the gentrification frontier (Holland 2012).

Ideally, we would take a longer-term view but, unfortunately, compatible census data is not available to catch the initial waves of gentrification in Islington and Notting Hill (*e.g.* Glass 1964), but we would expect any analysis of neighbourhood change in London using 2001–2011 data to pick up signs of status changes in areas such as London Fields, Dalston, Brixton and Peckham (Butler and Robson 2001; Benson and Jackson 2017). It might, of course, also show up changes associated with super-gentrification in neighbourhoods that

experienced gentrification in earlier periods (see Butler and Lees 2006 on Barnsbury), as well as areas demonstrating forms of incumbent improvement where displacement has not been a significant factor which is something that Freeman *et al*. (2015) suggest could well apply in London.

Scoring Results

Even after robust re-scaling, property prices and incomes 'count' for more than changes in skills or occupational mix in our scores, and following PCA the percent of variance explained by the first component (our score) is 78.8%. If we understand this as a way of mapping the data onto new axes aligned with variation in the 'data cloud', then the discarded components—accounting for 15.1%, 4.9% and 1.2% of variance respectively—capture lesser variation that we can loosely term 'noise' even though they might, in the round, still prove useful for prediction.

Error! Reference source not found. shows two axes of high property values emanating from Central London—Southwest and North-Northeast—with 'Billionaire's Row' (Bishop's Avenue) on Barnet's border with Haringey featuring prominently. In the context of an 'affordability crisis' in London housing (see Hamnett and Reades 2018), the emphasis on property price in our measurement of neighbourhood status encapsulates one of the main mechanisms through which even fairly well-off residents are experiencing neighbourhood change (Benson and Jackson 2017).

[Insert Figure 2 here]

Model Comparisons

Hyperparameter tuning—optimising for Mean Squared Error (MSE)—yielded a RF with a configuration of: 1,400 trees, 85% of features considered by each tree, no maximum tree depth, and a minimum leaf size of two. Compared to traditional methods (**Error!**

Reference source not found.), the RF shows improvements over both types of linear regression even without tuning, but the tuned model outperforms multiple linear regression by more than 10% across every measure.

[Insert Table 1 here]

However, the ultimate value of the model lies in how well it predicts the 2011 scores using the 2001 data: a Pearson's *r* of 0.99 indicates that for most observations the forest performs very well indeed. There *are* outliers of course, though it is reasonable to expect that major property developments, as well as the 'decanting' of residents from council estates undergoing redevelopment (*e.g.* Lees 2014), might transform individual neighbourhoods in ways that no predictive model could anticipate.

Predictor importance

Before introducing the predictions in detail we examine which variables the model found most important for predicting status change. A feature importance measure is automatically generated by RFs and is best understood as the contribution of the variable to the model. This metric is measured out of a theoretical maximum value of 1—so larger values mean more useful variables—but with 166 variables it is impractical to show these in a table and a visual representation has been used instead.

[Insert Figure 3 here]

is broadly consistent with hypotheses that relate to occupation and skills changes as drivers of neighbourhood change (Hamnett 2015): work-related variables make up much of the

top-20, with long hours (for both men and women), skills and qualifications (both high and low), and job flexibility (self-employment with and without employees, as well as homeworking) all good predictors of neighbourhood status change. Immigration from the Americas, 2001 EU members, and Oceania also show up in the top-30, suggesting that global-scale inflows are also a useful predictor (see Butler and Lees 2006). Older buildings remain attractive to in-movers (as hypothesised by Glass 1964 and many others), but rather less expected is the fact that 'DINKs' (Dual-Income, No Kids) do not feature strongly, though this is consistent with Karsten's (2003) observation of a shift towards child-rearing in the 'Inner City'.

[Insert Figure 3 here]

Trajectories of change

Taking an overview, **Error! Reference source not found.**a shows the changing distribution of scores over time, suggesting a flattening of the distribution whilst implying continued status change likely to have a pronounced impact on the most affordable and least-well off LSOAs. Note, however, that this trend is *not* expected to accelerate: **Error! Reference source not found.**b predicts an overall slowing of the magnitude of change. The neighbourhoods that have experienced the strongest change in 2001–2011 show comparably less change in the subsequent period.

[Insert Figure 4 here]

The more interesting analysis, however, is a geographical one: where is change most significant across the two-time periods? Since *everywhere* is experiencing status score

increase over the period 2001–2021 it is more useful to examine relative changes in the *ranking* of LSOAs. We could have random fluctuations in the rankings based on very minor differences in input variables, so it would be preferable to avoid taking 'noise' an indicator of significant change. Accordingly, since the distribution of changes in rank was broadly both symmetric and normal, these movements were grouped by standard deviation: more extreme values are more likely to indicate meaningful change. Movements within ±1 Standard Deviation are not shown in **Error! Reference source not found.** on the basis that they are most likely to represent random fluctuation.

[Insert Figure 5 here]

Broadly, **Error! Reference source not found.** shows Inner East London—those areas near the London Olympic development especially—'catching up' with *non-prime* West London. This is *not* to suggest that West London has seen some sort of decline, only that it is improving at a slower rate. 'Prime London' in Westminster and Kensington & Chelsea obviously saw enormous gains in 2001–2011, but the significant changes were concentrated towards the north ends of both boroughs where pockets of deprivation and un-upgraded housing remain.

Running the predictions forward to 2021 (**Error! Reference source not found.**) sees these concentrations disperse, though this should not be confused with an absence of change in these areas. What is striking about the comparison with **Error! Reference source not found.** is the shift outwards from Inner East London: a wedge of 'uplift' now extends out to the traditionally working class boroughs of Havering, Waltham Forest, and Bexley. 'Prime London' continues to pull away from the rest of the city in absolute terms, and we expect the vestiges of deprivation in these boroughs to be wiped out by the ongoing redevelopment

of council estates in both Westminster and Kensington & Chelsea (Lees 2014; Minton 2017).

In contrast, there are areas of relative decline in the outer boroughs of Croydon, Harrow and Hounslow implying that these are less likely to experience the changes and displacements associated with improving levels of education and in-movers engaged in higher-status work (see Leckie 2009 and Butler *et al*. 2013 on links between education and gentrification in London). A further implication is that the uplift of the East End may well be linked to displacement of the least well-off to Outer London (Travers *et al*. 2016)—something that Freeman *et al*. (2015:2811) also see as a distinct possibility given both that the poor are forced to move more frequently than the well-off, and that those moving into gentrifying areas are nearly three times more likely to have a degree than those moving into disadvantaged neighbourhoods.

[Insert Figure 6 here]

Discussion & limitations

For those who live in London, and who have the benefit of hindsight, some of these predictions may appear self-evident: these areas are on most people's radar and might even be seen to be areas where change has 'been and gone'. However, it is worth recognising that the *preconditions* of these changes must have been in place by 2011 for these predictions to be made and that, had we had access to this data *in* 2011, then we could have made these predictions at that time! It is therefore possible to envision revisions to our approach to incorporate more 'timely' data—such as from *Zoopla* (a property price website) or *Twitter* (useful as a marker of cultural change)—to develop the kind of real-time 'early warning system' anticipated by Chapple and Zuk (2016).

Although we have singled out Hamnett (2003) for his erroneous prediction of 'no change' in Clapton (Hackney) there is, of course, no guarantee that we will do better. Nonetheless,

if studies of gentrification and neighbourhood change are to offer more than a rigorous post-mortem then intensive case studies *must* be confronted with—and complemented by—predictions stemming from other approaches. Indeed, we hope to be proven wrong in some of our predictions, but explaining *why* we got it wrong should enrich understanding of the factors influencing areas in transition. For instance, Lees (2000:398) has noted there is a temporal aspect to change which means that the gentrifiers of today are not necessarily the same as those of the 1980s, so a clear limitation of the approach is that the model links the markers of change in 2011–2021 to those of 2001–2011. That said, it should also be recognised that the algorithm is not impacted by our human propensity to simplify and generalise, so while ML may be vulnerable to unforeseen behavioural change it is also more subtle in terms of how it makes use of the available data.

Regardless, longer-term data going back to 1981 or 1991 would benefit our approach substantially and enable us to explore the regeneration of the Docklands in the 1980s (Foster 1999) alongside trends highlighted by Hamnett (2009). Unfortunately, we have no equivalent to the US Neighbourhood Change Database (Barton 2016:7) which provides comparable data across multiple Censuses, and changes in the classification of account and small employers present additional challenges in using data of this vintage (Hamnett 2015:240–241). The absence of a gridded population surface on the Northern Irish model (*e.g.* Martin *et al.* 2011) also limits longitudinal research because of incompatible zone definitions; although the 'PopChange' project (Lloyd *et al.* 2016) is a promising step in this regard it is insufficient in terms of both resolution and the variables available.

Another factor that we have not directly addressed in this paper is the influence of neighbouring zones and 'edge effects': Redfern has argued that gentrification operates by a diffusion process (1997:1337), and Kolko (2007) noted that the income of adjacent census tracts might be a useful predictor of future neighbourhood change. It is likely that the incorporation of, for example, spatial lags via Local Indicators of Spatial Association (Anselin 1995) might improve our predictions. Moreover, change does not magically cease at the edge of London's administrative boundaries: we know that the past two decades have been characterised by the increasing suburbanisation of poverty (Travers *et al*. 2016) and

would have liked to expand our analysis beyond the GLA boundary but income data is not available at the LSOA scale outside of London.

There is, however, nothing to ultimately prevent us modelling the entire UK to search for larger patterns of neighbourhood change such as rural in-migration or the impact of empty second homes in areas such as Devon or Cornwall. Achieving this, however, will require the development of a deeper understanding of the typologies of neighbourhood change captured by the scoring metric through its interactions with the ML algorithm, something we anticipate undertaking as a piece of follow-on work in due course.

Conclusion

Gentrification research remains mired in debates about cause and effect, and whether displacement inevitably accompanies neighbourhood improvement (Hamnett 2003; Lees 2000; Freeman *et al*. 2015). Quantitative work has something to contribute here, showing where status change is occurring and relating it to other variables in a way that generates useful hypotheses about mechanisms of change. Not unlike qualitative work, such approaches also generate interesting, and at times counter-intuitive, findings about neighbourhood change (see, for example, Freeman *et al*.'s 2015 conclusion that there is no elevated mobility out of those London neighbourhoods experiencing gentrification).

However, in contrast to the quasi-experimental approach of Freeman *et al*. (2015) which said little about future trends, this paper has used innovative ML techniques to highlight neighbourhoods that are likely to significantly improve or decline by 2021. As well as noting the residualisation of some parts of outer London, our results suggest continuing 'uplift' in Inner East London and the spread of this process to the Outer Boroughs. Changes in neighbourhood status are, not unsurprisingly, strongly associated with house prices, the proportion of males and females in work for more than 30 hours a week, household incomes, and the share of knowledge workers, homeworkers, and professionals. It is these factors, as opposed to local amenities or travel, that appear worthy of more detailed exploration. That said, recent political developments, such as Brexit and changes to

London's infrastructure (*e.g*. Crossrail), mean that, while the specific predictions in this paper are unlikely to be accurate, they still provide a basis for further comparative investigation.

As a demonstration of the capabilities of Machine Learning in an urban studies context, this paper is a useful marker of the need for a rapprochement across the 'qualitative/quantitative divide'. We are not claiming to have explained or 'solved' the problem of neighbourhood change, nor are we suggesting that our approach supersedes the intensive, on-the-ground work undertaken by so many before, but it does open a new 'front' in our attempts to understand and, ultimately, anticipate neighbourhood transition. We hope that, in making these predictions about change in London, we are ultimately able to identify the ways that improvement or regeneration can occur without incurring displacement or disconcerting social change. Perhaps our predictions will be wrong for all the right reasons?

Acknowledgements

First, we would like to acknowledge the valuable contribution of Dr. Elizabeth Sklar, Jordan's co-supervisor on the original work that ultimately led to this article; her advice was integral to this research, though any errors or omissions remain ours alone. In addition, Jordan also wishes to acknowledge the contribution of Ivy Du to this work. We also made extensive use of the contributions of the many developers who have made possible Scikit-Learn 0.18 (Pedregosa *et al*. 2011) and Pandas (McKinney 2010) under version 3.6 of the Python programming language. The reproducible notebooks are made possible by Jupyter 4.1.0 (Kluyver 2016). The maps were created in QGIS 2.18 (Quantum GIS Development Team 2017). Other figures were produced in R using ggplot2 (Wickham 2009). All tools are available as Free Open Source Software. The codebase, including installation and configuration script for the required Python libraries, is available for download at: https://github.com/jreades/urb-studies-predicting-gentrification.

Appendix (production notes)

Our approach made intensive use of the Scikit-Learn toolkit 0.18 (Pedregosa *et al*. 2011) and Pandas (McKinney 2010) under version 3.6 of the Python programming language. The reproducible notebooks are made possible by Jupyter 4.1.0 (Kluyver 2016). The maps were created in QGIS 2.18 (Quantum GIS Development Team 2017). Other figures were produced in R using ggplot2 (Wickham 2009). All tools are available as Free Open Source Software. The GitHub repository is available at [not included to retain anonymity of peer review].

References

Anselin, L. (1995). Local Indicators of Spatial Association–LISA. *Geographical Analysis*, 27(2):93–115.

Arribas-Bel, D., P. Nijkamp and H. Scholten (2011). Multidimensional urban sprawl in Europe: A self-organizing map approach. *Computers, Environment and Urban Systems*, 35(4):263–275.

Arribas-Bel, D., J. Patino and J. Duque (2017). Remote sensing-based measurement of living environment deprivation: Improving classical approaches with machine learning. *PLoS ONE*, 12(5):e0176684.

Atkinson, R. (2000). Measuring gentrification and displacement in Greater London. *Urban Studies*, 37(1), 149–165.

Brunsdon, C. (2016). Quantitative methods I: Reproducible research and quantitative geography. *Progress in Human Geography*, 40(5):687–696.

Barton, M. (2016). An exploration of the importance of the strategy used to identify gentrification. *Urban Studies*, 53(1):92–111.

Benson, M. and E. Jackson (2017). Making the middle classes on shifting ground? Residential status, performativity and middle-class subjectivities in contemporary London British Journal of Sociology 68(2):215–233.

Bostic, R.W. and R.W. Martin (2003). Black home-owners as a gentrifying force? neighbourhood dynamics in the context of minority home-ownership. *Urban Studies*, 40(12):2427–2449.

Boy, J.D. and J. Uitermark (2016). How to study the city on Instagram. *PloS ONE*, 11(6):e0158161.

Breiman, L. (2001) Random forests. *Machine Learning*, 45(1):5–32.

Butler, T. and L. Lees (2006). Super-gentrification in Barnsbury, London: globalization and gentrifying global elites at the neighbourhood level. *Transactions of the Institute of British Geographers*, *31*(4), 467–487.

Butler, T. and G. Robson (2001). Social capital, gentrification and neighbourhood change in London: a comparison of three South London neighbourhoods. *Urban Studies*, 38(12):2145–2162.

Butler, T., C. Hamnett and M.J. Ramsden (2013). Gentrification, education and exclusionary displacement in East London. *International Journal of Urban and Regional Research*, 37(2), 556–575.

Chapple, K. (2009). *Mapping susceptibility to gentrification: The early warning toolkit*. Technical report, Centre for Community Innovation, University of California Berkeley, August 2009. URL: <https://communityinnovation.berkeley.edu/reports/Gentrification-Report.pdf>.

Chapple, K. and M. Zuk (2016). Forewarned: The use of neighborhood early warning systems for gentrification and displacement. *Cityscape: A Journal of Policy Development and Research*, 18(3):109–130.

Clark, E. (1988). The rent gap and transformation of the built environment: Case studies in Malmö 1860-1985. *Geografiska Annaler. Series B. Human Geography*, 241–254.

Cockings, S., A. Harfoot, D. Martin and D. Hornby (2011). Maintaining existing zoning systems using automated zone-design techniques: methods for creating the 2011 Census output geographies for England and Wales. *Environment and Planning A*, *43*(10), 2399–2418.

Cole, H.V., M.G. Lamarca, J.J. Connolly and I. Anguelovski (2017). Are green cities healthy and equitable? Unpacking the relationship between health, green space and gentrification. *J Epidemiol Community Health*, *71*(11), 1118–1121.

Dalton, C.M. and J. Thatcher (2015). Inflated granularity: Spatial 'Big Data' and geodemographics. *Big Data & Society*, 2(2):1–15.

Davidson, M. and L. Lees (2005). New-build 'gentrification' and London's riverside renaissance. *Environment and Planning A*, 37(7):1165–1190.

Donaldson, D. and A. Storeygard (2016). The view from above: Applications of satellite data in economics. *The Journal of Economic Perspectives*, 30(4):171–198.

Fan, C., S.J. Rey and S.W. Myint (2016, online early). Spatially filtered ridge regression (SFRR): A regression framework to understanding impacts of land cover patterns on urban climate. *Transactions in GIS*. doi: 10.1111/tgis.12240.

Foster, J. (1999). *Docklands: cultures in conflict, worlds in collision*. Psychology Press.

Freeman, L. (2005). Displacement or succession? Residential mobility in gentrifying neighborhoods. *Urban Affairs Review*, 40(4):463–491.

Freeman, L. (2009). Neighbourhood diversity, metropolitan segregation and gentrification: What are the links in the US? *Urban Studies*, 46(10):2079–2101.

Freeman, L., A. Cassola and T. Cai (2015). Displacement and gentrification in England and Wales: A quasi-experimental approach. *Urban Studies* 53(13):2797–2814.

Gale, C.G. (2014). Creating an open geodemographic classification using the UK Census of the Population. PhD thesis, University College London.

Gale, C.G. and P.A. Longley (2013). Temporal uncertainty in a small area open geodemographic classification. *Transactions in GIS*, 17(4):563–588.

Galster, G. (2001). On the nature of neighbourhood. *Urban Studies*, 38(12):2111–2124.

Glass, R.L. (1964). *London: aspects of change*. MacGibbon & Kee.

Guerts, P., D. Ernst and L. Wehenkel (2016). Extremely randomized trees. *Machine Learning*, 63(1):3-42.

Hagenauer, J. and M. Helbich (2017). A comparative study of machine learning classifiers for modeling travel mode choice. *Expert Systems with Applications*, 78:273–282.

Hamnett, C. (1983). Regional variations in house prices and house price inflation 1969–81. *Area*, 15(2): 97–109.

Hamnett, C. (1984). Gentrification and residential location theory: a review and assessment. *Geography and the Urban Environment: Progress in Research and Applications*, 6:283–319.

Hamnett, C. (2003). Gentrification and the middle-class remaking of inner London, 1961–2001. *Urban Studies*, 40(12):2401–2426.

Hamnett, C. (2009). Spatially displaced demand and the changing geography of house prices in London, 1995–2006. *Housing Studies*, 24(3):301–320.

Hamnett, C. (2015). The changing occupational class composition of London. *City*, 19(2–3):239–246.

Hamnett, C. and J. Reades (2018), Mind the Gap: implications of overseas investment for regional house price divergence in Britain. *Housing Studies*, doi: 10.1080/02673037.2018.1444151 .

Harris, A. (2012). Art and gentrification: pursuing the urban pastoral in Hoxton, London. *Transactions of the Institute of British Geographers*, 37(2), 226–241.

Hochstenbach, C., S. Musterd and A. Teernstra (2015). Gentrification in Amsterdam: Assessing the importance of context. *Population, Space and Place*, *21*(8), 754–770.

Holland, M. (2012) Chatsworth Road: the frontier of Hackney's gentrification, *The Guardian*, URL: <https://www.theguardian.com/uk/2012/jul/07/chatsworth-road-frontline-hackney-gentrification> [Last checked 17 February 2017]

Hristova, D., M. Williams, M. Musolesi, P. Panzarasa and C. Mascolo (2016) Measuring urban social diversity using interconnected geo-social networks. International World Wide Web Conference, Montreal, Canada, 2016; doi:10.1145/2872427.2883065.

Jackson, E. and M. Benson (2014). Neither 'Deepest, Darkest Peckham' nor 'Run-of-the-Mill' East Dulwich: The Middle Classes and their 'Others' in an Inner-London

Neighbourhood. *International Journal of Urban and Regional Research*, 38(4), 1195–1210.

James, G., D. Witten, T. Hastie and R. Tibshirani (2013). *An introduction to statistical learning*. London: Springer, 102, pp.303–368.

Karsten, L. (2003). Family gentrifiers: challenging the city as a place simultaneously to build a career and to raise children. *Urban Studies* 40(12):2573–2584.

Kluyver, T., B. Ragan-Kelley, F. Pérez, B. Granger, M. Bussonnier , J. Frederic, K. Kelley, J. Hamrick, J. Grout, S. Corlay, P. Ivanov, D. Avila, S. Abdalla, C. Willing and Jupyter Development Team (2016). Jupyter Notebooks—a publishing format for reproducible computational workflows. *Positioning and Power in Academic Publishing: Players, Agents and Agendas* (F. Loizides and B. Schmidt *eds*), IOS Press. URL <http://ebooks.iospress.nl/publication/42900> [Last checked 20 August 2017].

Kolko, J. (2007). The determinants of gentrification. *SSRN*, December 2007. doi: 10.2139/ssrn.985714.

Lansley, G. (2016). Cars and socio-economics: understanding neighbourhood variations in car characteristics from administrative data. *Regional Studies, Regional Science*, 3(1):264–285.

Lauria, M. and M.E. Stout (1995). The significance of scale in the analysis of gentrification. *College of Urban and Public Affairs (CUPA) Working Papers*, 1991–2000:9, University of New Orleans.

Leckie, G. (2009). The complexity of school and neighbourhood effects and movements of pupils on school differences in models of educational achievement. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 172(3), 537–554.

Lees, L. (2000). A reappraisal of gentrification: towards a 'geography of gentrification'. *Progress in Human Geography*, 24(3):389–408.

Lees, L. (2003). Super-gentrification: The case of Brooklyn Heights, New York City. *Urban Studies*, 40(12):2487–2509.

Lees, L. (2014). The urban injustices of new Labour's "New Urban Renewal": The case of the Aylesbury Estate in London. *Antipode*, *46*(4), 921–947.

Ley, D. (1986). Alternative explanations for inner-city gentrification: a Canadian assessment. *Annals of the Association of American Geographers*, 76(4), 521–535.

Li, Y. and Y. Xie (2018). A New Urban Typology Model Adapting Data Mining Analytics to Examine Dominant Trajectories of Neighborhood Change: A Case of Metro Detroit. *Annals of the American Association of Geographers*, doi: 10.1080/24694452.2018.1433016.

Liu, L., E.A. Silva, C. Wu and H. Wang (2017). A machine learning-based method for the large-scale evaluation of the qualities of the urban environment. *Computers, Environment and Urban Systems*, 65:113-125.

Lloyd, C.D., N. Bearman, G. Catney, A. Singleton and P. Williamson (2016) PopChange. Liverpool: Centre for Spatial Demographics Research, University of Liverpool. URL: <https://www.liverpool.ac.uk/geography-and-planning/research/popchange/introduction/>

Manley, D. and R. Johnston (2014). London: A dividing city, 2001–11? *City*, 18(6):633–643, 2014.

Manley, E., J. Addison and T. Cheng (2015). Shortest path or anchor-based route choice: a large-scale empirical analysis of minicab routing in London. *Journal of Transport Geography*, 43:123–139.

Martin, D., C. Lloyd and I. Shuttleworth (2011). Evaluation of gridded population models using 2001 Northern Ireland Census data. *Environment & Planning A*, 43(8):1965–1980.

Mavrommatis, G. (2011). Stories from Brixton: Gentrification and Different Differences. *Sociological Research Online*, 16(2):1–10.

McKinney, W. (2010). Data Structures for Statistical Computing in Python, *Proceedings of the 9th Python in Science Conference*, 51–56.

Melchert, D. and J.L. Naroff (1987). Central city revitalization: A predictive model. *Real Estate Economics*, 15(1):664–683.

Meligrana, J. and A. Skaburskis (2005). Extent, location and profiles of continuing gentrification in Canadian metropolitan areas, 1981-2001. *Urban Studies*, 42(9):1569–1592.

Minton, A. (2017) *Big Capital: who is London for?* Penguin.

Morenoff, J.D. and M. Tienda (1997). Underclass neighborhoods in temporal and ecological perspective. *The Annals of the American Academy of Political and Social Science*, 551(1):59–72.

Naik, N., S.D. Kominers, R. Raskar, E.L. Glaeser and C.A. Hidalgo (2017). Computer vision uncovers predictors of physical urban change. *Proceedings of the National Academy of Sciences*, 114(29):7571–7576.

Neal, S., G. Mohan, A. Cochrane and K. Bennett (2016). 'You can't move in Hackney without bumping into an anthropologist': why certain places attract research attention. *Qualitative Research*, 16(5):491–507.

O'Sullivan, D. (2002). Toward micro-scale spatial modeling of gentrification. *Journal of Geographical Systems*, *4*(3), 251-274.

Office for National Statistics (nd). *Census geography*. URL: <https://www.ons.gov.uk/methodology/ geography/ukgeographies/censusgeography - super-output-area-soa> [Last checked 17 August 2017].

Owens, A. (2012) Neighborhoods on the rise: A typology of neighborhoods experiencing socioeconomic ascent. *City & Community*, 11(4):345–369.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot and E. Duchesnay (2011). Scikit-learn: Machine Learning in Python, *JMLR* 12:2825–2830.

Quantum GIS Development Team (2017). Quantum GIS Geographic Information System. *Open Source Geospatial Foundation Project*. URL: <http://qgis.osgeo.org/>.

Reades, J. and D. Smith. (2014) Mapping the 'Space of Flows': the geography of global business telecommunications and employment specialisation in the London Mega-City Region. *Regional Studies*, 48(1):105–126.

Reades, J. (2014) Mapping changes in the affordability of London with open-source software and open data: 1997–2012. *Regional Studies, Regional Science*, 1(1):336–338.

Redfern, P.A. (1997). A new look at gentrification: 2. A model of gentrification. *Environment and Planning A*, 29(8):1335–1354.

Redfern, P.A. (2003). What makes gentrification 'gentrification'? *Urban Studies*, 40(12):2351–2366.

Rey, S.J. (2014). Rank-based Markov chains for regional income distribution dynamics. *Journal of Geographical Systems*, 16(2):115–137.

Santibanez, S.F., M. Kloft and T. Lakes (2015). Performance Analysis of Machine Learning Algorithms for Regression of Spatial Variables: A Case Study in the Real Estate Industry. Paper presented to *GeoComputation*, Dallas, pp.292–297.

Singleton, A.D., S. Spielman and C. Brunsdon (2016). Establishing a framework for Open Geographic Information science. *International Journal of Geographical Information Science*, 30(8):1507–1521.

Slater, T. (2009). Missing Marcuse: On gentrification and displacement. *City*, 13(2–3):292–311.

Smith, N. (1996). *The new urban frontier: Gentrification and the revanchist city*. Psychology Press.

Steif, K., A. Mallac, M. Fichman and S. Kassel (2017). *Predicting gentrification using longitudinal census dat*a. URL: <http://urbanspatialanalysis.com/portfolio/predicting-gentrification-using-longitudinal-census-data/> [Last checked 17 August 2017].

Stevens, F.R., A.E. Gaughan, C. Linard and A.J. Tatem (2015). Disaggregating census data for population mapping using random forests with remotely-sensed and ancillary data. *PloS One*, 10(2):e0107042.

Sturgis, P., I. Brunton-Smith, J. Kuha and J. Jackson (2014). Ethnic diversity, segregation and the social cohesion of neighbourhoods in London. *Ethnic and Racial Studies*, *37*(8):1286–1309.

Travers, T., S. Sims and N. Bosetti (2016). *Housing and inequality in London*. Technical report, Centre for London.

van Criekingen, M. and J.-M. Decroly (2003). Revisiting the diversity of gentrification: neighbourhood renewal processes in Brussels and Montreal. *Urban Studies*, 40(12):2451–2468.

van Ham M., D. Manley, N. Bailey, L. Simpson, D. Maclennan (2012) Neighbourhood Effects Research: New Perspectives. In: van Ham M., Manley D., Bailey N., Simpson L., Maclennan D. (eds) *Neighbourhood Effects Research: New Perspectives*, 1–21. Springer, Dordrecht.

Vickers, D. and P. Rees (2007). Creating the UK national statistics 2001 output area classification. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 170(2):379–403.

Watt, P. (2008). 'The only class in town? Gentrification and the middle-class colonization of the city and the urban imagination'. *International Journal of Urban and Regional Research*, 32:206–211.

Watt, P. (2013). 'It's not for us': Regeneration, the 2012 Olympics and the gentrification of East London. *City*, 17(1):99–118.

Wickham, H. (2009). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag, New York. URL: <http://ggplot2.org/>.

Wyly, E. (2014). Automated (post)positivism. *Urban Geography*, 35(5):669–690.

Xiao, N. (2016). Machine Learning in Richardson, D. (ed) *Encyclopaedia of Human Geography*. John Wiley & Sons, Ltd, doi: 10.1002/9781118786352.wbieg0673.

Zhong, C., S.M. Arisona, X. Huang, M. Batty and G. Schmitt (2014). Detecting the dynamics of urban structure through spatial network analysis. *International Journal of Geographical Information Science*, 28(11):2178–2199.

Zukin, S., V. Trujillo, P. Frase, D. Jackson, T. Recuber and A. Walker (2009). New retail capital and neighborhood change: boutiques and gentrification in New York City. *City & Community*, 8(1):47–64.

Zukin, S., S. Lindeman and L. Hurson (2017) The omnivore's neighborhood? Online restaurant reviews, race, and gentrification. *Journal of Consumer Culture*, 17 (3):459–479.

Table 1. Model Comparison

Model	R^2	Expl. Var.	MSE	MAE
Simple Linear Regression[1]	0.528	0.538	0.294	0.343
Multiple Linear Regression[2]	0.639	0.640	0.225	0.305
Extremely Random Trees (Default)	0.649	0.653	0.219	0.284
Extremely Random Trees (Tuned)	0.699	0.703	0.188	0.259

[1] Using the strongest predictor variable (median house prices).

[2] Using all 166 variables.

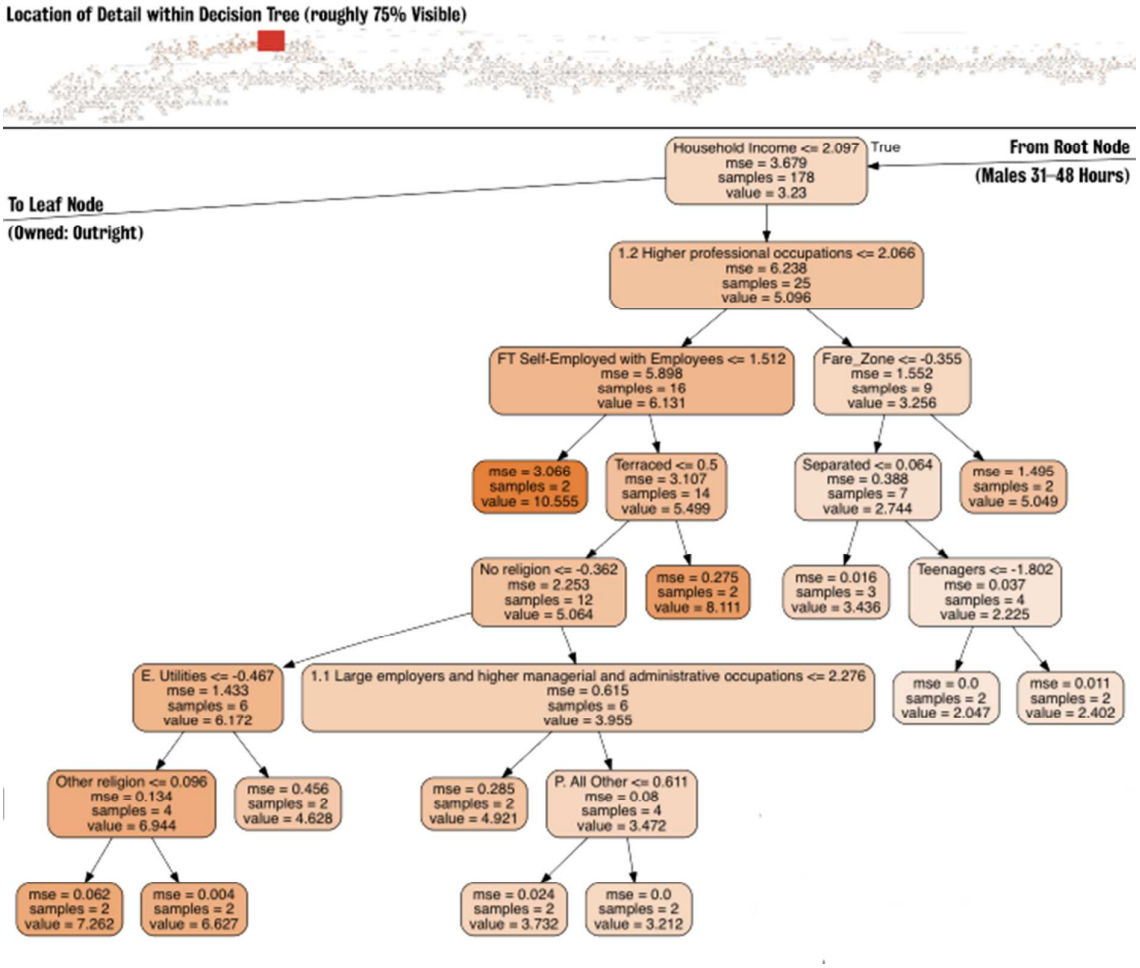

Figure 1. Detail from a regression tree used by the Random Forest in this research[1]

[1] Each leaf node shows: the variable and value used in the split; the Mean Squared Error of the prediction for all observations in this region; the number of observations (samples); and the predicted value for observations in this region (this will usually be the mean).

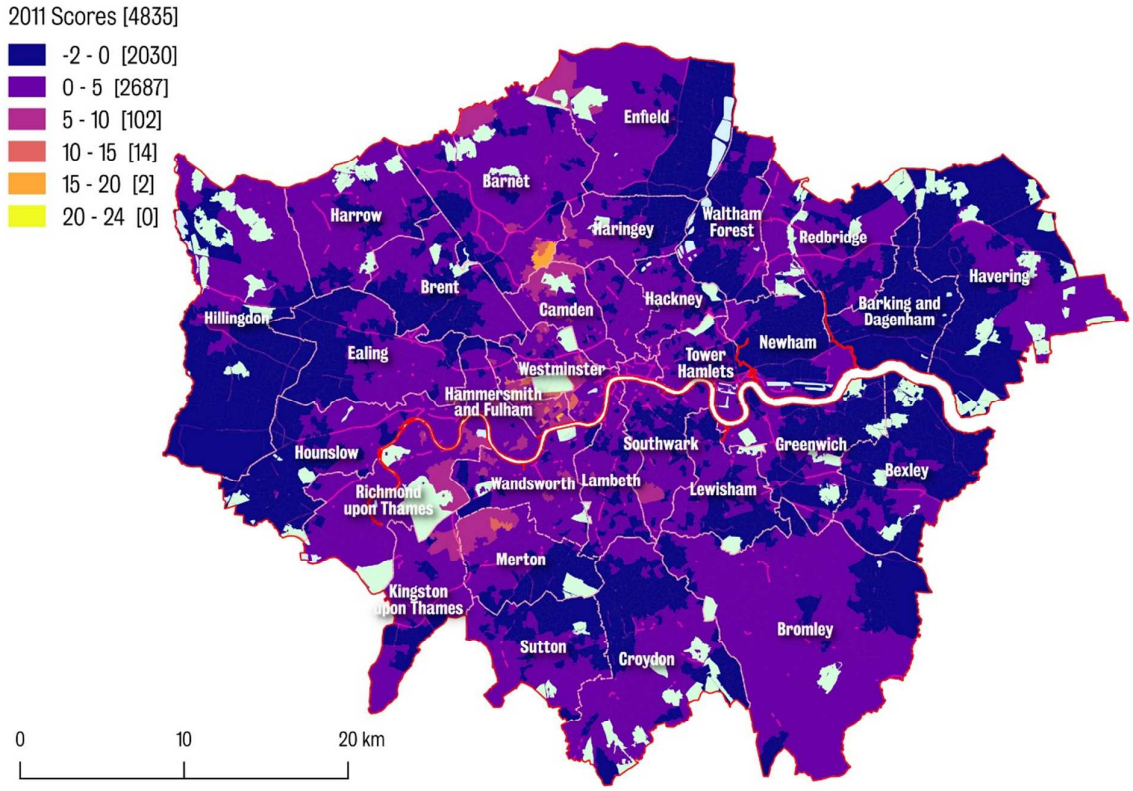

Figure 2. 2011 Status Scores for LSOAs

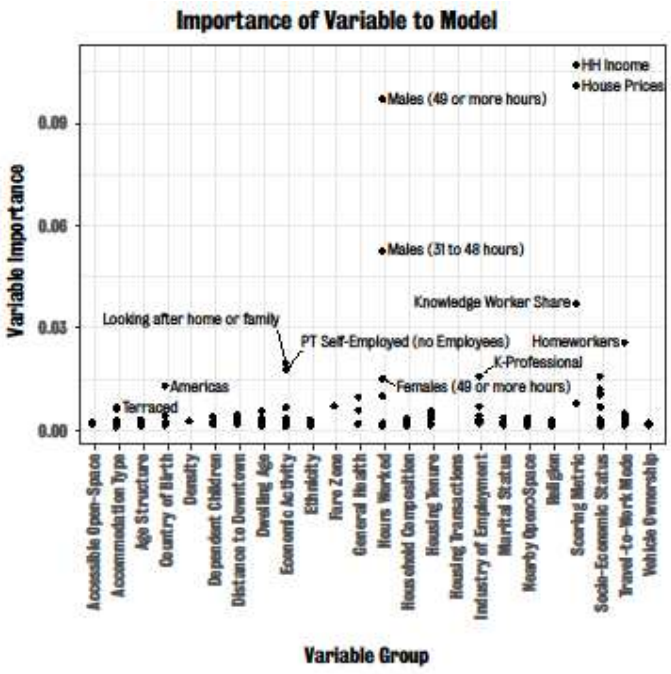

Figure 3. Parameter Importance to Tuned Model (Grouped by Variable Category)

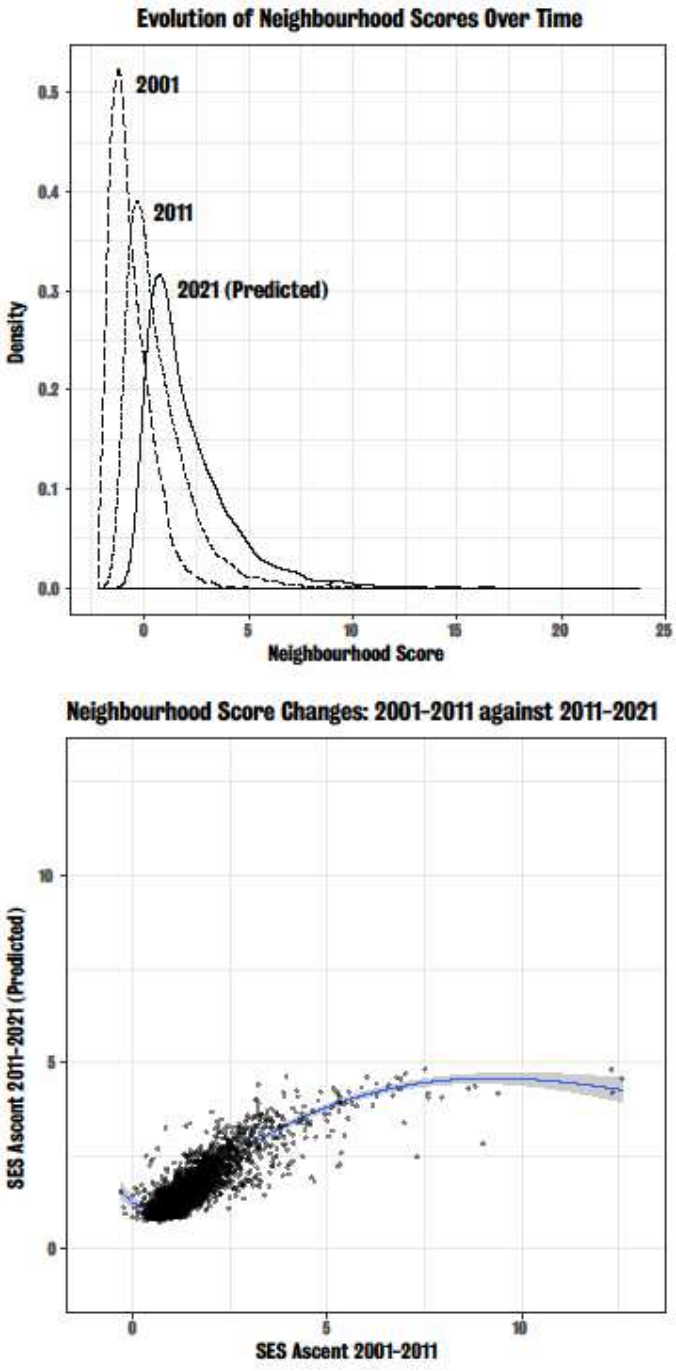

Figure 4a and b. Score Change Over Time

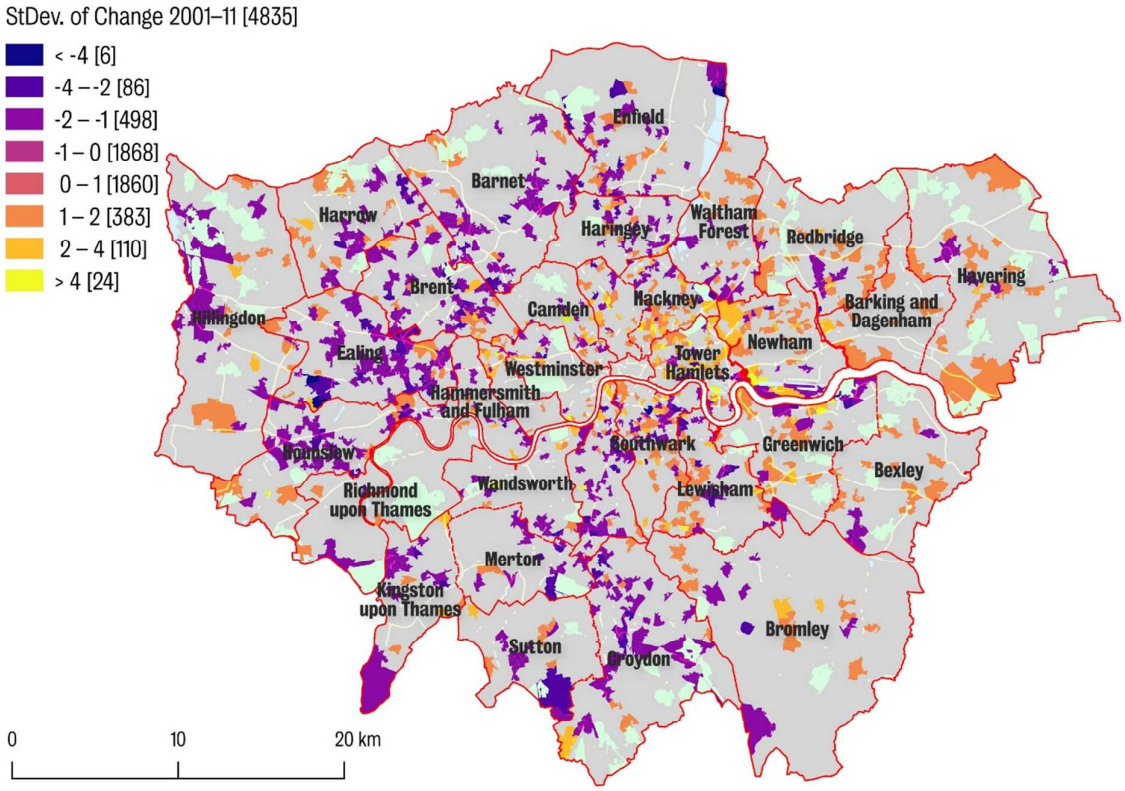

Figure 5. Standard Deviation of Change in Rank 2001–2011 (±1 not shown)

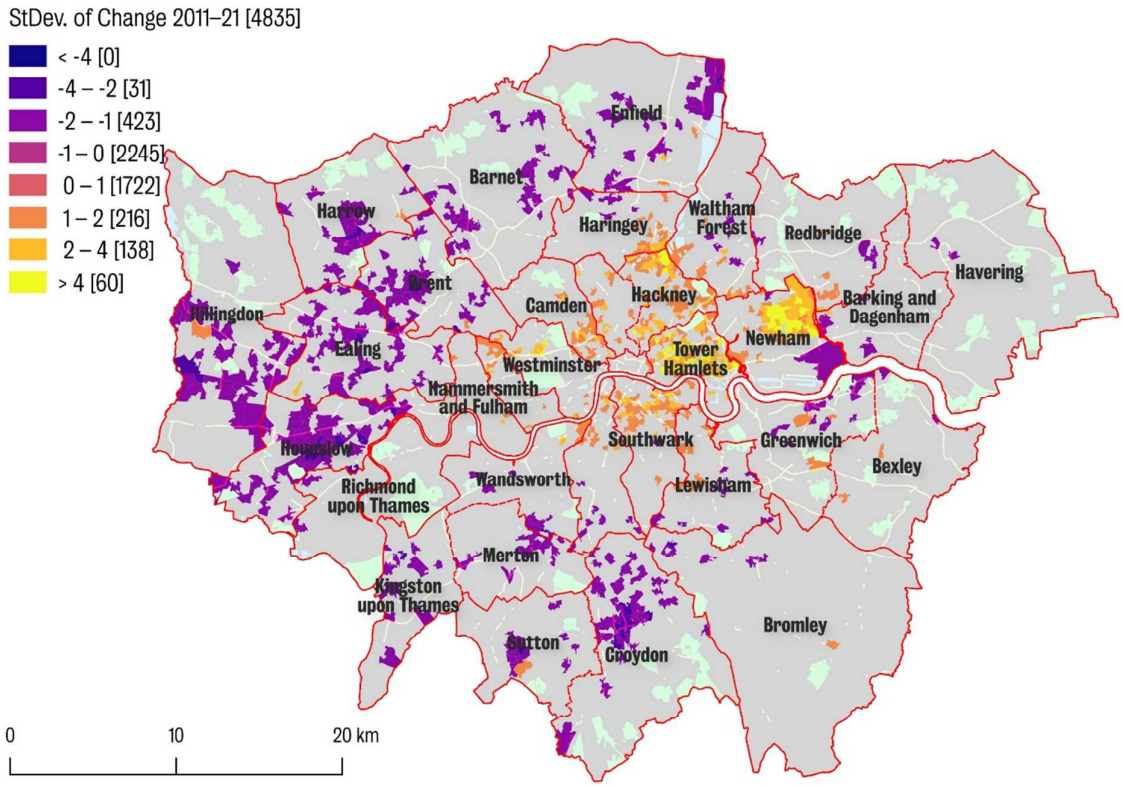

Figure 6. Standard Deviation of Change in Rank 2011–2021 (±1 not shown)